



# TODD WARCZAK, Ph.D.

## Data Scientist

### CONTACT

twarczak@gmail.com 

(206) 999-6478 

White River Junction, VT 

[toddwarczak.netlify.app](https://toddwarczak.netlify.app) 

[LinkedIn](#)  [Github](#)  [@TWarczak](#) 

### EDUCATION

#### Ph.D.

Molecular & Cellular Biology

Dartmouth College

September 2012 - August 2020

#### B.S.

Biology - University of Utah

September 2008 - May 2012

### SKILLS

#### Coding

R ★★★★★, Python ★★, SQL ★★★

#### AWS Cloud

SageMaker, S3, EC2, ...

#### Reports/Collaboration

Github, Markdown, Jupyter

#### Data Cleaning/Wrangling

tidyverse ★★★★★, dplyr, purrr,

lubridate, stringr, tidyr, ...

#### Custom/Automated Data Viz

ggplot2 ★★★★★, Plotly,

Leaflet/ggmaps/sf (geospatial),

modeltime (auto-time-series), ...

#### Machine Learning

tidymodels, SageMaker, recipes

TensorFlow, Keras, H2O

#### Time-Series Forecasting

modeltime, lubridate, timetk

#### General

Probability & Statistics, NLP

Microsoft Excel/PowerP/Word,

Regression/Classification/Clustering

### WORK EXPERIENCE

#### Data Scientist / Molecular Biologist

Dartmouth College, September 2012 - September 2020 / Hanover, NH

- Engineered novel genome-wide association study (GWAS) that identified genes controlling arsenic tolerance in plant roots. Utilized expert data cleaning, wrangling, and analytic skills to summarise findings from millions of observations across thousands of unique genomes.
- Determined plant gene AtNIP1;1 is the major genetic factor for tolerating arsenic in root cells and identified regions of interest on multiple chromosomes.
- Built lab RNA-seq pipeline for gene expression of 25000+ plant genes with R scripts for gene clustering (PCA, hierarchical clustering), regression (GLMs/ANOVA), exploratory data analysis, and causal inference.

### PROJECTS

#### SageMaker + RStudio to Predict Home Prices w/ Multi-class XGBoost; Explaining Model Behavior with Geospatial Plots and SHAP

- Explored Austin dataset to predict home price in [Kaggle](#) competition. ([Blog](#), [Github](#))
- Built static and interactive geospatial plots overlaid with feature data.
- Feature engineered high/low important words that associate w/ price using NLP
- Trained/tuned/evaluated/deployed SageMaker Multi-class XGBoost on holdout data & submitted predicted binned price to Kaggle competition. Submission scored 0.8876 (mLogLoss), which would have placed 6th (out of 90 entries) in live competition.
- Modified {SHAPforxgboost} package to generate multi-class SHapley Additive exPlanations (SHAP) values/plots that explain how XGBoost model made predictions.

#### Predicting Churn using AWS SageMaker & Local RStudio

- Leveraged SageMaker and AWS tools to train/tune/evaluate/deploy XGBoost model for predicting bank customer churn in SLICED [Kaggle](#) competition. ([Blog](#), [Github](#))
- Configured local RStudio to make API calls to SageMaker using SageMaker Python SDK and {reticulate}.
- Top model deployed as SageMaker endpoint for real-time predictions on holdout data (w/ minimal feature engineering and pre-processing).
- Predictions submitted to SLICED competition received a score of 0.07622 (LogLoss), which would have placed 8th out of 130 entries.

#### Forecasting Daily Sales with {modeltime}

- Performed EDA and generated 3 month forecasts of daily sales for 'Superstore' company ([Kaggle](#)) selling furniture, technology, and office supplies. ([Blog](#), [Github](#))
- Tested multiple {tidymodels} workflows with {modeltime} for time-series forecasting.
- Best individual models combined to generate single weighted ensemble forecast (Support Vector Machine/Neural Network-AR/Random Forest/Prophet-XGboost).

#### TidyTuesday

- Weekly twitter project focusing on cleaning, wrangling, summarizing, and arranging a new dataset in R to produce a single chart. Typically using {ggplot2} and {tidyverse} tools. Shared via #TidyTuesday. ([Github](#))